Bias Mitigation in Big Data Systems: A Systematic Review, Comparative Synthesis, and Decision Framework for Fair Algorithm Design

Daniel Kweku Assumang[1*], Ifeoma Eleweke[2], Chiamaka Ezenwaka[3], Olabode Soetan[4], Martin Msughter Vincent[5], Elizabeth Ayodeji Adeyefa[6]
[1]College of Professional Studies, Roux Institute, Northeastern University, Maine USA
[2]College of Technology and Engineering, Westcliff University, California USA
[3]Department of Operation, CBRE Group Inc, Texas USA
[4]Department of Tax Technology Consulting practice, Deloitte LLP, Chicago USA
[5]National Graduate Institute for Policy Studies
[6]School of Business, Economics and Technology, Campbellsville University, USA

E-mail: dkassumang96@gmail.com

Keywords: Bias mitigation, big data, algorithmic fairness, fair algorithm design, ethical AI, automated decision-making

Received: July 28, 2025

As Big Data systems increasingly shape decision-making across healthcare, finance, hiring, criminal justice, and public governance, concerns regarding algorithmic bias and unfair outcomes have become more pronounced. While data-driven systems can improve efficiency, prediction, and service delivery, they may also reproduce historical inequalities through biased datasets, proxy variables, flawed labels, or opaque optimization processes.

This study presents a systematic review and comparative synthesis of bias mitigation strategies in Big Data systems, with the aim of identifying effective approaches for fair algorithm design. Guided by a PRISMA-informed review methodology, literature published between 2018 and 2026 was screened across major academic databases, resulting in 27 studies included for final analysis.

The findings classify mitigation approaches into pre-processing, in-processing, post-processing, and hybrid methods. Pre-processing techniques were found effective for addressing representation imbalance and data quality issues, while in-processing methods provided stronger fairness control where model retraining was feasible. Post-processing approaches were most practical for legacy or proprietary systems, whereas hybrid strategies were strongest in high-risk contexts requiring layered safeguards. The review further shows that no single fairness metric or mitigation technique is universally optimal; effectiveness depends on domain risk, bias source, regulatory obligations, and operational constraints.

Based on these findings, the paper proposes the Fair Algorithm Design Decision Framework to guide organizations in selecting context-appropriate fairness interventions. The study concludes that bias mitigation should be treated as a continuous lifecycle responsibility integrating data governance, model design, human oversight, and ongoing monitoring.

Povzetek: Študija primerja pristope za zmanjševanje algoritmične pristranskosti ter poudarja, da je treba pravičnost zagotavljati celostno, neprekinjeno in glede na kontekst uporabe.

1 Introduction

Background to the study

Big Data systems increasingly shape decision-making across healthcare, finance, criminal justice, hiring, marketing, and public governance. The growing volume, velocity, and variety of data allow organizations to improve forecasting, optimize operations, personalize services, and support real-time decision processes (Hossin et al., 2023; Aljehani et al., 2024).

Despite these benefits, Big Data systems can reproduce and amplify inequality when bias enters the data pipeline or model design. Bias may arise from unrepresentative sampling, historical discrimination embedded in legacy records, subjective labeling, proxy variables, or feedback loops after deployment. Prior studies have documented

biased outcomes in recruitment systems, criminal risk scoring, healthcare prediction, and credit assessment (Mittelstadt et al., 2016; Chen, 2023; Shahul Hameed et al., 2024).

Accordingly, the challenge is not only to build accurate predictive systems, but also to design systems that are fair, transparent, accountable, and socially responsible. As the use of automated decision-making expands, reducing bias has become a central requirement for trustworthy Big Data systems (Shahul Hameed et al., 2024).

Problem statement

A central challenge in Big Data systems is that bias may arise at multiple stages, including data collection, labeling, feature selection, model training, and deployment. Common forms include sampling bias, where certain populations are underrepresented, and historical bias, where past inequalities are encoded into training data (Mittelstadt et al., 2016). Once embedded in machine learning systems, these biases can scale rapidly and disproportionately disadvantage women, minority groups, and economically vulnerable populations (Chohlas-Wood et al., 2023).

In recruitment, algorithms trained on historical hiring decisions may replicate gender bias, particularly in male-dominated sectors (Chen, 2023; Ho et al., 2025; Köchling and Wehner, 2020). In healthcare, poorly designed models may underperform for minority populations if demographic variation is not adequately represented in training data (Shahul Hameed et al., 2024). In criminal justice, biased risk assessment tools may reinforce racial disparities (Joseph, 2024).

These examples demonstrate that algorithmic bias is not merely a technical error; it is a governance and social justice issue requiring robust mitigation strategies and careful system design.

Research gap

Although algorithmic fairness has received growing scholarly attention, much of the existing literature either discusses ethical principles at a conceptual level or evaluates individual mitigation techniques in isolation. Comparative evidence across pre-processing, in-processing, post-processing, and hybrid strategies remains fragmented. Furthermore, limited guidance exists on how organizations should choose appropriate mitigation methods under varying data conditions, fairness objectives, and operational constraints.

Research questions

This study addresses the following research questions:

RQ1: What are the principal sources of bias in Big Data systems?

RQ2: How do pre-processing, in-processing, post-processing, and hybrid mitigation methods compare in effectiveness, scalability, and fairness–accuracy trade-offs?

RQ3: How do mitigation approaches vary across domains such as healthcare, finance, hiring, and criminal justice?

RQ4: What decision framework can guide fair algorithm design in practice?

Contributions

This paper makes three contributions. First, it presents a systematic review of recent literature on bias mitigation in Big Data systems. Second, it provides a comparative synthesis of major mitigation approaches across technical and application dimensions. Third, it proposes a decision framework to support the practical selection of fairness strategies based on domain risk, data quality, and deployment constraints.

2 Methodology

2.1 Review design and search strategy

This study employed a PRISMA-guided systematic literature review to identify and synthesize evidence on bias mitigation in Big Data systems. Searches were conducted across Google Scholar, IEEE Xplore, ScienceDirect, Scopus, Web of Science, and PubMed using combinations of terms related to algorithmic fairness, bias mitigation, ethical AI, healthcare bias, hiring bias, and automated decision-making. The review focused on studies published between 2018 and 2026.

2.2 Study selection and eligibility

The initial search identified 200 records, including 14 studies obtained through backward citation screening. After removal of 38 duplicates, 162 studies underwent title and abstract screening, resulting in 61 full-text articles assessed for eligibility. Following exclusion of studies lacking methodological rigor, fairness relevance, or sufficient analytical value, 27 studies were retained for comparative synthesis.

Studies were included if they examined bias in Big Data or automated decision systems, evaluated mitigation approaches or fairness metrics, and were published in peer-reviewed English-language sources. Editorials, inaccessible full texts, and studies unrelated to fairness or algorithmic bias were excluded. The PRISMA selection process is summarized in Table 1 and Figure 1.

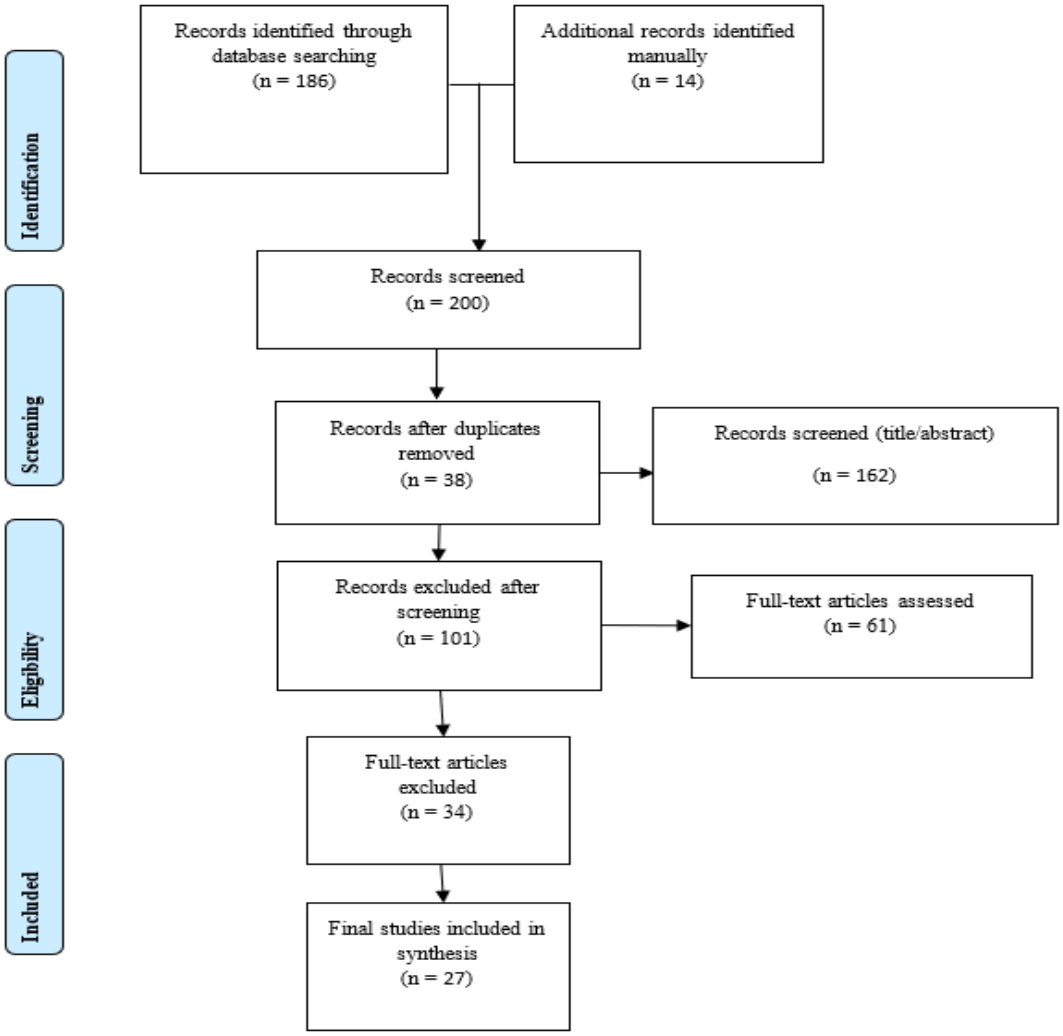
2.3 Data extraction and comparative synthesis

A structured extraction framework was used to record study characteristics, including application domain, bias category, mitigation stage, fairness objective, methodological design, and reported outcomes. Included studies were coded into four mitigation categories: pre-processing, in-processing, post-processing, and hybrid/governance approaches.

Because reviewed studies differed substantially in datasets, metrics, and evaluation methods, a comparative narrative synthesis was adopted rather than statistical meta-analysis. Comparative analysis focused on fairness improvement, predictive accuracy, interpretability, scalability, and suitability for high-risk deployment contexts. Greater analytical weight was assigned to empirical and systematic review studies with transparent methodology and measurable fairness outcomes.

Table 1: PRISMA Flow diagram

PRISMA Stage	Count
Records identified through database search	186
Additional records identified manually	14
Total records identified	200
Duplicates removed	38
Records screened (title/abstract)	162
Records excluded after screening	101
Full-text articles assessed	61
Full-text articles excluded	34
Final studies included in synthesis	27



3 Foundations of bias in big data systems

Bias in Big Data systems refers to systematic distortions in data collection, representation, model learning, or decision outputs that produce unfair outcomes across individuals or groups. In algorithmic environments, bias may emerge from historical inequalities, incomplete data capture, proxy variables, measurement error, optimization objectives, or deployment feedback effects rather than explicit malicious intent (Mittelstadt et al., 2016; Ferrara, 2023). Because automated systems operate at scale, such distortions can influence large populations and significantly affect access to employment, healthcare, credit, and public services.

Table 2: Major sources of bias in big data systems

Bias Type	Stage	Example	Typical Risk	Author
Sampling Bias	Data collection	Underrepresented populations	Poor subgroup accuracy	Chen et al., 2021; Shahul Hameed et al., 2024;
Historical Bias	Legacy datasets	Past hiring inequality	Reproduced discrimination	Hanna et al., 2024; Varsha, 2023
Label Bias	Annotation	Subjective recruiter evaluations	Skewed learning	Favier et al., 2023; Ferrara,

				2023; Hanna et al., 2024
Measurement Bias	Feature capture	Unequal diagnostics	Hidden unfairness	Chohlas-Wood et al., 2023

Bias may therefore emerge at multiple stages of the machine learning lifecycle, including data collection, feature engineering, model optimization, and deployment. Even where protected attributes are removed, correlated variables may continue to reproduce discriminatory outcomes (Wang et al., 2024). Similarly, models trained on historical records may inherit social or institutional inequalities embedded within legacy datasets (Köchling and Wehner, 2020; Joseph, 2024). Deployment environments may further intensify unfairness through feedback loops, automation overreliance, and changing population behavior.

Big Data environments amplify these risks because scale, velocity, and interconnected data ecosystems increase both the reach and persistence of biased decisions. Automated outputs may propagate across downstream systems such as lending, insurance pricing, hiring, and public administration, while reduced transparency can make fairness failures difficult to detect or correct (Ukanwa, 2024).

These findings indicate that bias mitigation must address multiple stages of the data lifecycle rather than relying on a single technical correction. This provides the foundation for the comparative review of mitigation strategies presented in Section 5.

4 Fairness definitions and evaluation metrics

Fairness in algorithmic decision-making refers to the principle that automated systems should not systematically disadvantage individuals or groups based on protected or socially sensitive characteristics such as race, gender, ethnicity, disability, or socio-economic status (Mittelstadt et al., 2016; Ferrara, 2023). However, fairness is not a single measurable property. Different contexts may prioritize equality of outcomes, equal opportunity, calibration, transparency, or procedural fairness, and these objectives may conflict when group base rates differ. Consequently, fairness evaluation must be aligned with domain-specific risks and operational goals rather than assuming a universally optimal metric.

Table 3: Major fairness metrics in algorithmic systems

Metric	Definition	Common Use	Key Limitation
Demographic Parity	Equal positive outcome rates across groups	Hiring, outreach systems	May ignore qualification differences
Equal Opportunity	Equal true positive rates	Healthcare, lending	Does not equalize false positives
Equalized Odds	Equal true and false positive rates	Criminal justice	Difficult to optimize jointly
Calibration	Comparable risk reliability across groups	Finance, insurance	May conflict with parity metrics
Individual Fairness	Similar individuals receive similar outcomes	Admissions, recruitment	Requires defining similarity
Counterfactual Fairness	Outcome unchanged under protected-attribute change	Causal fairness analysis	Complex causal assumptions

Fairness metrics also involve important trade-offs. Strict parity constraints may reduce aggregate predictive accuracy, while calibration objectives may conflict with equalized outcome measures when base rates differ. Existing metrics further simplify fairness into statistical relationships and may overlook structural inequality, intersectional harms, or procedural legitimacy (Hagendorff, 2020; Singh et al., 2024). Consequently, fairness metrics should be treated as decision-support tools rather than complete ethical solutions.

Because mitigation techniques optimize different fairness objectives, selecting an intervention without first defining the relevant fairness metric risks addressing the wrong problem. This provides the basis for the comparative review of mitigation strategies presented in Section 5.

5 Comparative review of bias mitigation techniques

Bias mitigation approaches are commonly categorized according to the stage of the machine learning lifecycle at which intervention occurs: pre-processing, in-processing, post-processing, and hybrid approaches. Each category addresses different sources of unfairness and involves trade-offs in scalability, interpretability, implementation complexity, and predictive performance (Ferrara, 2023; Varsha, 2023). No single mitigation strategy is universally optimal, as effectiveness depends on the dominant source of bias, fairness objective, deployment environment, and regulatory context.

Table 4: Comparative summary of bias mitigation techniques

Category	Typical Methods	Main Strengths	Main Limitations	Best Context	Author
Pre-processing	Reweighing, resampling, synthetic data generation	Scalable, model-agnostic, improves representation	May not address latent proxy bias	Imbalanced or structured datasets	Holstein et al., 2023; Hosseinzadeh et al., 2021; Qi et al., 2019; Almasoud and Idowu, 2025
In-processing	Fairness constraints, adversarial debiasing, fair representation learning	Strong fairness control during training	Technically complex, less interpretable	Custom ML pipelines	Shahul Hameed et al., 2024; Chohlas-Wood et al., 2023
Post-processing	Threshold adjustment, reject option classification, output calibration	Fast deployment, no retraining required	Often corrective rather than preventive	Legacy or vendor-controlled systems	Hanna et al., 2023; Mittelstadt et al. 2016
Hybrid	Multi-stage mitigation and governance integration	Broad protection across lifecycle stages	Higher cost and governance burden	High-risk domains	Holstein et al., 2023; Hosseinzadeh et al., 2021; Qi et al., 2019; Almasoud and Idowu, 2025; Belenguer, 2022

Pre-processing methods are particularly effective where bias arises from sampling imbalance, historical underrepresentation, or noisy labels (Holstein et al., 2023; Hosseinzadeh et al., 2021; Qi et al., 2019). In-processing approaches provide stronger fairness control by embedding constraints directly into model optimization, although they often require advanced technical expertise and reduced interpretability (Shahul Hameed et al., 2024; Chohlas-Wood et al., 2023). Post-processing methods are operationally attractive when models cannot easily be retrained, while hybrid approaches are increasingly favored in high-risk domains because they address multiple bias sources simultaneously (Singh et al., 2024; Farayola et al., 2025). Recent studies also highlight the growing importance of adversarial debiasing and fair representation learning in healthcare and other high-dimensional AI environments (Hasanzadeh et al., 2025).

Across studies, three recurring trade-offs emerge: fairness versus predictive accuracy, fairness versus interpretability, and fairness versus operational cost. Strict parity constraints may reduce aggregate performance where group base rates differ, while advanced debiasing models may become difficult to explain to regulators or affected users. Similarly, layered monitoring and governance mechanisms improve fairness resilience but increase implementation complexity and resource requirements. These findings suggest that fairness should be managed as a strategic lifecycle objective rather than treated as an isolated technical adjustment.

The comparative evidence further indicates that mitigation effectiveness depends less on identifying a universally

superior algorithm and more on matching interventions to specific bias sources and operational conditions. This provides the basis for the domain-specific findings presented in Section 6.

6 Domain-specific findings

Although fairness principles are theoretically generalizable, the expression of algorithmic bias varies substantially across domains due to differences in data quality, regulatory expectations, decision risk, historical inequality, and operational constraints. The comparative synthesis indicates that domain context strongly influences dominant bias sources, fairness metric selection, acceptable accuracy trade-offs, and governance requirements.

Table 5: Cross-domain comparative summary

Domain	Primary Harm Risk	Common Bias Sources	Relevant Metrics	Typical Mitigation Priorities
Healthcare	Missed care / unsafe outcomes	Underrepresentation, measurement bias	Equal opportunity, calibration	Subgroup validation, balanced data, clinician oversight
Finance	Exclusion / unlawful discrimination	Proxy bias, historical exclusion	Calibration, disparate impact	Explainability, audits, proxy testing
Hiring	Opportunity denial	Historical labels, proxy variables	Demographic parity, adverse impact	Ranking audits, human review
Criminal Justice	Liberty and trust harms	Historical enforcement, feedback loops	Equalized odds, calibration	Transparency, public oversight
Public Governance	Welfare exclusion / legitimacy	Incomplete records, access gaps	Inclusion metrics, procedural fairness	Appeals processes, governance controls

Healthcare systems frequently experience fairness challenges related to demographic underrepresentation and measurement bias, particularly in diagnostic and predictive models (Hanna et al., 2024; Hasanzadeh et al., 2025). Financial systems are more strongly affected by proxy discrimination and regulatory transparency concerns, while hiring systems often inherit historical recruitment inequalities through biased labels and ranking patterns (Chen, 2023; Köchling and Wehner, 2020). In criminal justice and public governance contexts, fairness concerns extend beyond statistical parity to include procedural legitimacy, public accountability, and the social consequences of automated decision-making (Joseph, 2024; Ukanwa, 2024).

Across domains, three findings consistently emerge. First, fairness metric selection is context-dependent rather than universal. Second, governance mechanisms such as transparency, appeals processes, and human oversight are often as important as technical mitigation methods. Third, data quality remains foundational, as poor representation and historical distortions frequently undermine fairness before model training begins.

The synthesis further suggests that fairness controls should scale proportionally with deployment risk. High-risk systems such as healthcare triage, criminal justice scoring, credit access, and welfare allocation require rigorous subgroup testing, documented fairness metrics, human oversight, and periodic auditing. Lower-risk systems may justify lighter but still meaningful monitoring approaches. This risk-tiered logic aligns with emerging governance frameworks such as the European Union AI Act (Kusche, 2024).

These findings directly inform the decision framework proposed in Section 7, which translates comparative evidence and domain risk into practical guidance for fair algorithm design.

7 Proposed decision framework for fair algorithm design

The comparative findings demonstrate that no single mitigation strategy is universally optimal. Fairness outcomes depend on bias source, data quality, model transparency, domain risk, and operational constraints. To address this challenge, this paper proposes the Fair

Algorithm Design Decision Framework (FADDF), a structured model for selecting, deploying, and monitoring fairness interventions in Big Data systems.

7.1 Framework logic

The framework supports five core decisions:

1. Define decision context and deployment risk
2. Diagnose dominant bias sources
3. Select appropriate fairness metrics
4. Match mitigation strategies to technical constraints
5. Establish continuous monitoring and governance controls

Table 6: Core Inputs to the Framework

Input Dimension	Key Question
Decision Risk	Can errors affect health, liberty, employment, finance, or essential services?
Bias Source	Is bias driven by imbalance, proxy variables, historical discrimination, or deployment drift?
Data Quality	Are groups sufficiently represented and labels reliable?
Model Access	Can the model be retrained or audited internally?
Explainability Need	Must outputs be interpretable to regulators or users?
Regulatory Exposure	Are there legal obligations regarding transparency or discrimination?

Table 7: Mitigation selection matrix

Condition	Preferred Strategy	Rationale
Imbalanced data	Pre-processing	Corrects representation before training
Proxy discrimination	In-processing	Better control over learned relationships
Vendor-controlled models	Post-processing	No retraining required
High-risk deployment	Hybrid	Multi-layer protection justified
Dynamic environments	Monitoring + retraining	Fairness may drift over time

7.2 Risk-tier governance model

The framework distinguishes between high-risk, moderate-risk, and lower-risk systems.

- High-risk systems (healthcare triage, criminal justice scoring, credit access, welfare eligibility) require documented fairness metrics, human oversight, external audits, appeals processes, and incident response procedures.
- Moderate-risk systems (fraud detection, workforce analytics, pricing optimization) require internal audits, threshold review, and periodic retraining assessment.
- Lower-risk systems (content recommendation and non-essential personalization) may justify lighter but still meaningful bias monitoring.

This proportional governance logic aligns with emerging regulatory approaches such as the European Union AI Act (Kusche, 2024).

Table 8: Example framework applications

Scenario	Metric	Mitigation	Governance
Hiring platform	Disparate impact	Reweighing + ranking audit	Appeals + adverse impact testing
Clinical risk model	Equal opportunity	Resampling + subgroup validation	Drift monitoring
Loan approval model	Calibration	Feature audit + constrained optimization	Regulatory audit trail

7.3 Implementation principles

Successful deployment of fairness controls requires organizations to:

1. diagnose bias before selecting interventions
2. align metrics with domain-specific harms
3. match controls to deployment risk
4. maintain human oversight in high-impact systems
5. monitor fairness continuously after deployment
6. document decisions for accountability and auditability

The proposed framework contributes to the literature by integrating bias diagnosis, metric selection, mitigation matching, lifecycle monitoring, and governance controls into a single operational model bridging fairness theory and organizational implementation.

8 Discussion

This review demonstrates that bias mitigation in Big Data systems cannot be reduced to a single technical solution or universal fairness metric. The effectiveness of fairness interventions depends on the interaction between data quality, model design, deployment context, governance structure, and domain-specific risk. Consequently, organizations should move beyond searching for a universally "fair" algorithm and instead adopt context-sensitive mitigation strategies aligned with operational realities and potential harms.

One of the clearest findings across reviewed studies is that fairness is highly domain-dependent. In healthcare, minimizing missed diagnoses and ensuring subgroup safety may take priority over equal outcome rates, whereas hiring systems emphasize adverse impact reduction and procedural consistency. Financial systems prioritize calibration, explainability, and auditability, while criminal justice systems require balanced error distribution and public legitimacy. These differences indicate that fairness metric selection is not purely technical, but also institutional and normative in nature.

The review further suggests that many fairness failures originate before model training begins. Underrepresentation, historical distortions, weak labels, and proxy variables frequently undermine fairness at the data level, regardless of model sophistication. This finding implies that organizations may overemphasize advanced debiasing algorithms while underinvesting in data governance, subgroup representation, documentation, and continuous dataset evaluation. Fairness also cannot be sustained through one-time interventions alone, as real-world systems are subject to deployment drift, changing populations, feedback loops, and evolving regulatory expectations. Effective governance therefore requires ongoing monitoring, human oversight, transparency mechanisms, and periodic reassessment of subgroup outcomes.

Recent research in adaptive and intelligent control systems further highlights the limitations of static bias mitigation approaches in dynamic and safety-critical environments (Elrifaee, & Zayed, 2025). In autonomous systems, robotics, and nonlinear control architectures, operational conditions may evolve continuously due to uncertainty, feedback interactions, sensor noise, or environmental drift (Haq, Felicetti, Carnı, & Lamonaca, 2026). Under such conditions, fairness interventions designed using static datasets or fixed optimization assumptions may degrade over time or fail to respond adequately to changing subgroup behavior. Adaptive and feedback-driven control paradigms, including fuzzy control and dynamic optimization frameworks, therefore suggest important directions for future fairness research by enabling continuous adjustment, real-time monitoring, and resilience under uncertain deployment conditions. These findings reinforce the importance of lifecycle-oriented fairness governance in high-risk AI systems operating under real-time or safety-critical constraints.

Several unresolved challenges remain. Current fairness metrics often inadequately address intersectional harms, fairness under limited data conditions, and the governance of large-scale generative AI systems. In addition, tensions persist between fairness, privacy, interpretability, and predictive accuracy, particularly in high-risk domains. These challenges reinforce the need for interdisciplinary research integrating machine learning, governance, law, ethics, and public policy in the development of trustworthy AI systems.

9 Limitations and future research

This review has several limitations. The included studies span multiple disciplines, datasets, model architectures, fairness metrics, and evaluation approaches, limiting direct comparability and making statistical meta-analysis inappropriate. In addition, fairness itself remains conceptually contested, with studies prioritizing demographic parity, equal opportunity, calibration, explainability, or broader equity objectives differently. Publication bias may also affect the evidence base, as studies reporting successful mitigation outcomes are more

likely to appear in academic literature than null or unsuccessful implementations. Furthermore, because the review focused primarily on English-language and peer-reviewed sources, relevant industry practices and non-English contributions may be underrepresented.

The broader fairness literature also faces several structural limitations. Many mitigation techniques are evaluated using static benchmark datasets rather than dynamic real-world deployment environments, limiting understanding of fairness drift, feedback effects, and long-term governance challenges. Existing research further provides limited evidence regarding implementation cost, organizational accountability structures, and the operational sustainability of fairness interventions across sectors.

Several priorities therefore emerge for future research. Increasing attention is needed for fairness monitoring in dynamic environments, particularly in relation to deployment drift, subgroup degradation, and continuous auditing systems. Emerging generative AI and foundation models also introduce new fairness risks involving representation harms, synthetic content generation, and downstream automation effects. Additional research is needed on intersectional fairness, fairness–privacy trade-offs, human-AI interaction, and cross-jurisdiction regulatory alignment in globally deployed systems.

The proposed Fair Algorithm Design Decision Framework should therefore be viewed as an adaptable operational model rather than a universal solution. Future empirical work should validate the framework through case studies, deployment experiments, and cross-sector implementation analysis in areas such as hiring, healthcare, finance, and public governance.

10 Conclusion

This study examined bias mitigation in Big Data systems through a systematic review, comparative synthesis, and framework-development perspective. The findings demonstrate that algorithmic fairness cannot be achieved through isolated technical corrections alone, as unfair outcomes often emerge from interactions between data quality, historical inequality, model design, deployment environments, and institutional governance. The review further showed that no single fairness metric or mitigation strategy is universally optimal. Instead, effective bias mitigation depends on aligning fairness objectives, technical methods, and governance controls with domain-specific risks and operational conditions.

To address this challenge, the study proposed the Fair Algorithm Design Decision Framework (FADDF), which integrates bias diagnosis, fairness metric selection, mitigation matching, risk-tier governance, and post-deployment monitoring into a unified operational model. The framework contributes practical guidance for organizations seeking to implement fairness controls across healthcare, finance, hiring, criminal justice, and public governance systems.

Ultimately, fair algorithm design should be understood as a continuous lifecycle responsibility rather than a one-time compliance exercise. As automated decision systems continue to expand across high-impact sectors, future progress will depend on stronger data governance, continuous auditing, interdisciplinary collaboration, and accountable AI oversight capable of balancing predictive performance with social legitimacy and public trust.

References

[1] Aljehani, S. B., Abdo, K. W., Alam, M. N., & Aloufi, E. M. (2024) ‘Big Data Analytics and Organizational Performance: Mediating roles of green innovation and knowledge management in telecommunications’, *Sustainability*, *16*(18), 7887. Available at: https://doi.org/10.3390/su16187887

[2] Almasoud, A. S., & Idowu, J. A. (2025). Algorithmic fairness in predictive policing. *AI and Ethics*, *5*(3), 2323-2337.

[3] Belenguer, L. (2022) ‘AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine centric solutions adapted from the pharmaceutical industry’, *AI Ethics*, 2(4), pp.771–787. Available at: https://doi.org/10.1007/s43681-022-00138-8

[4] Chen, S., Keglovits, M., Devine, M., & Stark, S. (2021) ‘Sociodemographic differences in respondent preferences for survey formats: sampling bias and potential threats to external validity’, *Archives of Rehabilitation Research and Clinical Translation*, *4*(1), 100175. Available at: https://doi.org/10.1016/j.arrct.2021.100175

[5] Chen, Z. (2023) ‘Ethics and discrimination in artificial intelligence-enabled recruitment practices’, *Humanities and Social Sciences Communications*, *10*(1). Available at: https://doi.org/10.1057/s41599-023-02079-x

[6] Chohlas-Wood, A., Coots, M., Goel, S. and Nyarko, J. (2023) ‘Designing equitable algorithms’, *Nature Computational Science*, Perspective. Available at: https://doi.org/10.1038/s43588-023-00485-4

[7] Elrifaee, M., & Zayed, T. (2025). Smart IoT-BIM framework with modified zonal safety analysis (ZSA) for real-time safety monitoring in construction. *Automation in Construction*, *178*, 106431.

[8] Farayola, M. M., Tal, I., Saber, T., Connolly, R., & Bendechache, M. (2025). A fairness-focused approach to recidivism prediction: implications for accuracy, trust, and equity. *AI & Society*. https://doi.org/10.1007/s00146-025-02452-1

[9] Favier, M., Calders, T., Pinxteren, S., & Meyer, J. (2023) 'How to be fair? A study of label and selection bias', *Machine Learning*, *112*(12), 5081–5104. https://doi.org/10.1007/s10994-023-06401-1

[10] Ferrara, E. (2023) 'Fairness and Bias in Artificial intelligence: A brief survey of sources, impacts, and mitigation strategies', *Sci*, *6*(1), 3. Available at: https://doi.org/10.3390/sci6010003

[11] Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

[12] Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. (2024) 'Ethical and bias considerations in artificial intelligence (AI)/Machine learning', *Modern Pathology*, 100686. Available at: https://doi.org/10.1016/j.modpat.2024.100686

[13] Haq, I. U., Felicetti, F., Carnı, D. L., & Lamonaca, F. (2026). Advancements in robotic positioning: A comprehensive review of measurement methods and systems. *Measurement: Sensors*, 102000.

[14] Hasanzadeh, F., Azizi, Z., White, J.A., Josephson, C.B. and Waters, G. (2025) 'Bias recognition and mitigation strategies in artificial intelligence healthcare applications', *npj Digital Medicine*, 8(154). Available at: https://doi.org/10.1038/s41746-025-01503-7

[15] Ho, J. Q., Hartanto, A., Koh, A., & Majeed, N. M. (2025) 'Gender Biases within Artificial Intelligence and ChatGPT: Evidence, Sources of Biases and Solutions', *Computers in Human Behavior Artificial Humans*, 100145. Available at: https://doi.org/10.1016/j.chbah.2025.100145

[16] Hosseinzadeh, M., Azhir, E., Ahmed, O.H. and Ghafour, M.Y. (2021) 'Data cleansing mechanisms and approaches for big data analytics: a systematic study', *Journal of Ambient Intelligence and Humanized Computing*, 14(4), pp.1–13. Available at: https://doi.org/10.1007/s12652-021-03590-2

[17] Hossin, M.A., Du, J., Mu, L. and Asante, I.O. (2023) 'Big Data-Driven Public Policy Decisions: Transformation toward smart Governance', SAGE Open, 13(4). Available at: https://journals.sagepub.com/doi/full/10.1177/215

[18] Joseph, J. Predicting crime or perpetuating bias? The AI dilemma. *AI & Soc* **40**, 2319–2321 (2025). https://doi.org/10.1007/s00146-024-02032-9

[19] Köchling, A., & Wehner, M. C. (2020) 'Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development', *BuR - Business Research*, *13*(3), 795–848. Available at: https://doi.org/10.1007/s40685-020-00134-w

[20] Kusche, I. (2024) 'Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk', *Journal of Risk Research*, 27(6). Available at: https://doi.org/10.1080/13669877.2024.2350720

[21] Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) 'The ethics of algorithms: Mapping the debate', *Big Data & Society*, July–December, pp. 1–21. Available at: https://doi.org/10.1177/2053951716679679

[22] Shahul Hameed, M.A., Qureshi, A.M. and Kaushik, A. (2024) 'Bias mitigation via synthetic data generation: A review', *Electronics*, 13(3909). Available at: https://doi.org/10.3390/electronics13193909

[23] Singh, N., Kapoor, A., & Soni, N. (2024). A sociotechnical perspective for explicit unfairness mitigation techniques for algorithm fairness. *International Journal of Information Management Data Insights*, *4*(2), 100259. https://doi.org/10.1016/j.jjimei.2024.100259

[24] Ukanwa, K. (2024). Algorithmic bias: Social science research integration through the 3-D Dependable AI Framework. *Current Opinion in Psychology*, *58*, 101836. Available at: https://doi.org/10.1016/j.copsyc.2024.101836

[25] Varsha, P.S. (2023) 'How can we manage biases in artificial intelligence systems – A systematic literature review', *International Journal of Information Management Data Insights*, 3(1), Article 100165. https://doi.org/10.1016/j.jjimei.2023.100165

[26] Wang, X., Wu, Y. C., Ji, X., & Fu, H. (2024). Algorithmic discrimination: examining its types and regulatory measures with emphasis on US legal practices. *Frontiers in Artificial Intelligence*, *7*. Available at: https://doi.org/10.3389/frai.2024.1320277